Estimating the Number of Concepts

Gregory Grefenstette

8
Estimating the Number of Concepts

Gregory Grefenstette

Abstract

Most Natural Language Processing systems have been built around the idea of a word being something found between white spaces and punctuation. This is a normal and efficient way to proceed. Tasks such as Word Sense Disambiguation, Machine Translation, or even indexing rarely go beyond the single word. Language models used in NLP applications are built on the word, with a few multiword expressions taken as exceptions. But future NLP systems will necessarily venture out into the uncharted areas of multiword expressions. The dimensions and the topology of multiword concepts are unknown: Are there hundreds of thousands or tens of millions? Which words participate in multiword concepts and which do not? As the corpus grows, will their number keep on increasing? In this paper, I estimate the number of multiword concepts that are used in English, systematically probing the Web as our corpus.

1. Introduction

It may seem futile to try to estimate how many concepts[1] there are in a language. Since man can create new concepts at will, one of his conceits as a sentient being, counting concepts can seem as silly as trying to determine how many numbers there are. Everyone knows that one can continue counting forever.

This viewpoint would be held by those who have worked closest with the lexicon. Patrick Hanks, esteemed editor of the *Encyclopedic World Dictionary* (Hanks & Potter 1971), the *Collins Dictionary of the English Language* (Hanks

et al. 1979), the *Oxford English Reference Dictionary* (Hanks *et al*. 1995), among others, has written:

> It is impossible to say how many words there are in the English language, because writers and speakers constantly create new terms to suit their purposes. (Hanks 2005: 248)

Kenneth Sisam, former Assistant Secretary at OUP, would concur, having written:

> I am afraid it is entirely impossible for us – or indeed for anyone – to say how many words there are in the English language. It depends on how many compounds etc. are counted and what are admitted as words. In the great Oxford English Dictionary there are said to be roundly 500,000 words defined. I am sorry I cannot be more precise. (Sisam 1930)[2]

Indeed, unbridled creativity is one of the hopes that moves civilisation forward, and this creativity finds its expression in new concepts and new uses of language.

But if one cannot bound the number of concepts, one can examine what concepts are in current use, and attempt to count these. I make a first pass on such an estimation here. In doing so, I will make a number of no doubt egregious approximations, many of which will provoke a frustrating mixture of ire and disdain from traditional lexicographers. But at least, I will start the ball rolling.

1.1. From words to concepts

The reason why I think such an attempt is needed is to prod computational lexicography and computational linguistics into the new millennium. Until recently, until the computer era, dictionaries and lexicography were essentially a paper-bound affair. Each new definition, each described concept, cost space on paper.

A dictionary was constrained by a budget, which limited the number of pages to be printed, which limited the number of words and concepts that could be included. Computational lexicons used for parsing have often been derived from printed dictionaries (Guthrie *et al*. 1996). Large hand-built computational lexical resources such as WordNet were also inspired by printed dictionaries (Miller *et al*. 1990).[3] Since the priority in dictionaries was to cover as many single words as possible in the limited space given, the decision to include a

given multiword expression was problematic, a decision often based on non-transparency of meaning or frequency.

This essentially word-based approach to language led to the development of computational parsers also based on words. Individual words (space-separated tokens) are often ambiguous, both syntactically and semantically. This problem has led to much research in computational linguistics, from part of speech tagging to Word Sense Disambiguation.[4] The standard technique for disambiguating a word is to start with a manually tagged text, each word tagged with its proper part of speech or sense. Then, using standard machine learning techniques,[5] creating a statistical model of the context of each disambiguated word that can be applied to unlabeled text.

All these attempts start from the word to the structure. But I think it is a safe bet that computational linguistics in the 21st century will work from larger structures than individual words and that these larger structures will no longer be seen as a *pain*[6] but as the normal structure that is manipulated and modelled by computational linguists.

Bank is an ambiguous word. *River bank* is not.

Computational lexicographers and linguists of the future will describe one-word or multiword concepts, no longer limited by space requirements. How much work will they have cut out for them? How many concepts will they have to model? We aim to provide the first estimate of this number in the next sections.

2. Concepts and two-word noun phrases

What is a concept? It has been explained as a mental representation, an abstract object, a cognitive unit of meaning (Margolis & Laurence 2007). Unfortunately, none of these explanations provides an operational definition.

We will start from a simpler definition and state that any noun is a concept. Verbs can also be concepts, of course, but in English, there is an order of magnitude more nouns than verbs.[7] If we find the number of concepts involving just nouns (and adjectives) we should not be very far from a good estimate.

Using the same, possibly faulty, reasoning, we will also ignore any concepts composed of three or more words. One justification for this last simplification of our task is the fact that most terms used in both folksonomies (Bibsonomy, CiteULike, Connotea) and edited thesauri (the Academic Computing Machinery subject listing, the Agriculture Information and Standards ontology, BioLinks, the thesaurus of the European Environment Information and Observation Net-

work, and Medical Subject Headings) are composed of one and two-word terms (cf. Good & Tennis 2009).

So if we can estimate the number of two-word phrases in common use, we can get an idea of how many concepts future computational linguists will have to model.[8]

2.1. A lower bound

Oxford University Press, on its AskOxford website, declares:

> The Second Edition of the *Oxford English Dictionary* contains full entries for 171,476 words in current use, and 47,156 obsolete words. To this may be added around 9,500 derivative words included as subentries. Over half of these words are nouns, about a quarter adjectives, and about a seventh verbs; the rest is made up of interjections, conjunctions, prepositions, suffixes, etc.[9]

If we take this number of 90 thousand nouns from OUP, and suppose that each noun corresponds to a concept, our lower bound is 90,000 concepts.

2.2. An upper bound

If we take each of the 90,000 nouns and suppose that they can be modified by another noun, e.g. *fire engine*, or by one of the 50,000 adjectives, then we have 90,000 x 140,000 ≈ 12 billion possibilities.

The question now becomes, between 90,000 and 12 billion, how many two-word terms are in current use?

2.3. Methodology

In order to estimate the number of current two-word noun phrases, I will use a list of nouns and adjectives from a downloadable dictionary, and a copy of the index of the Web.

3. Probing the Web

At Exalead, we crawl and index the Web. We hold 8 billion URLs to Web pages in our index at the moment, corresponding to about 4 billion Web pages crawled and indexed. The other 4 billion URLs are found in these Web pages and have not yet been crawled, but they can be indexed by using the text point-

ing to them. 4 billion pages is a subset of the entire Web. The estimates of the number of pages available on the Web run over 50 billion.

If we take this (large) sample of 4 billion pages, we can probe its index and find out which words are found together. The simplest method would be to take all the nouns and all the adjectives of English and to look for all two-word combinations.

To specify what we mean by adjective and noun, we began with the DELA dictionaries (downloaded in October 2008). From this lexicon, we extracted all unique surface forms tagged as nouns and containing only lowercase letters. There are 160,966 such nouns (from *aah*, *aahs*, *aalii*, *aaliis*, *aandblom*, *aandbloms*, *aardvark*, *aardvarks*, … through … *zymoses*, *zymosimeter*, *zymosimeters*, *zymosis*, *zymotechnics*, *zymurgies*, *zymurgy*, *zythum*. Similarly for adjectives, we find 46,759 adjectives (from *abactinal*, *abandoned*, *abapical*, *abased*, *abatable*, *abatised*, *abattoirlike*, *abaxial*, *abbatial*, … through … *zygoneurous*, *zygophyllaceous*, *zygose*, *zygosporic*, *zygotic*, *zymogenic*, *zymologic*, *zymological*, *zymolytic*, *zymotic*. If we generated all adjective-noun and noun-noun pairs, we would have over 7.5 billion combinations.

Some combinations appear often on the Web. For example, *mad dog* appears on over 400,000 Web pages (out of 8 billion). Some never appear. For example, there is no Web page with the phrase *abactinal zymosis*.[10] If the present paper were to be indexed someday on the Web, we would then find the hitherto-unknown concept of *abactinal zymosis*. In this way the formerly nonsensical phrase *colorless green ideas*, that Noam Chomsky invented for *Syntactic Structures* in 1957 has found its way onto over 4,000 Web pages in our index.[11]

In order to estimate *real* concepts, whatever that might mean, we decided to only consider two-word combinations that appeared on over 5 different Web pages. This threshold is chosen without any justification whatsoever, except that it would eliminate the indexed version of this article (and the four other articles that would cite the *abactinal zymosis* example from this text).

From another partial crawl of the Web, we have the word counts of the 2 million most-frequent strings. In that crawl, we find the string *the* 1.77 billion times, which, in other words, allows us to estimate the total number of English words crawled to be about 100 billion. In this large sample of English there are some of the DELA words that are never found. If we take only those nouns that appear more than 5 times, we are left with 89,257 nouns, from *aah*, *aahs*, *aalii*, *aaliis*, *aardvark*, *aardvarks*, *aardwolf*, *aardwolves*, … through … *zygote*, *zygotene*, *zygotes*, *zymase*, *zymogen*, *zymogens*, *zymology*, *zymolysis*, *zymosis*, *zymurgy*. Similar filtering with the adjective list leaves us with 25,687 adjectives, from *abactinal*, *abandoned*, *abased*, *abatable*, *abaxial*, *abbatial*, *abbrevi-*

ate, *abdicable*, *abdominal*, … through … *zoophilic*, *zootechnical*, *zwitterionic*, *zygodactyl*, *zygomatic*, *zygomorphic*, *zygomorphous*, *zygotic*, *zymolytic*, *zymotic*. We still have the potential pair *abactinal zymosis*, but now we are down to (89,257 + 25,687) x 89,257, or a little over 10 billion combinations to check.

Most Web search engines will allow you to make up to one query per second without blacklisting your computer. We could test all the possible two-word combinations, but at 1 query per second, it would still take 325 years. To speed things up, we decided to generate random samples of adjective-noun and noun-noun pairings from the attested nouns and adjectives, and to check their frequency in our Exalead online Web index. Here are some examples: *enhydra encyclopedists*, *cylindric chelicerate*, *radionics convoker*, *zoophile cognitions*, *nonrenewal cheerlessness*, *infidel kyanites*, *nondrying dormitories*, *languorous cyclohexanol*, *sawbones mortars*, *hedges braais*.[12]

Over 1 million such pairs were generated, covering 0.01% of the possibilities. Each pair was used as a contiguous query (the two words had to appear next to each other on the same page). Of these million possibilities, about 61,000 were attested in the index of 4 billion Web pages. About 21 thousand appeared on only one page, 8,663 appeared on 2 pages, 4,932 appeared on 3, etc. Table 1 shows a breakdown of the types of pairs found.

3.1. Estimate

If we take this sample, and consider those appearing on over 10 Web pages (out of 4 billion) as in common use, then we find 18,396, or about 1.8% of the generated pairs. If we extrapolate this to the entire sample of 10 billion combinations, this leaves us with about 180 million two-word combinations in 'common use' on the Web. This is an extremely rough estimate, and probably more of an upperbound than a true estimate of actual two-word noun phrases that are commonly used. But it gives a first idea of the scale that computational lexicography will have to deal with in the future.

3.2. Improvements

As anyone can see, even in the small samples of pairs listed, there is much room for improvement.

A number of suggestions for improving the counts follow, revolving around: Language identification, Not ignoring punctuation, Part of Speech tagging, Shallow parsing, Two-word phrases, Proper Name recognition, Ignoring spam, and Starting from the most common words.

Table 1: Breakdown of the types of two-word pairs.

attested on pages	*number of pairs*	*percent generated*	*sample pairs*
0	977,161	97.2%	reallocation trination, aestheticians gully, jumpier jerkin, unillustrated spanks, pastoral intemperateness, donax directives, hoofprints hygromas, teleworks pronoun, grout mas, cooptation geometries
1	20,950	2%	dusky chum, inevitable starling, sharpshooter planes, excipients abusing, comparative practising, spelt thundering, reserves onions, rotgut rotter, differences chroniclers, catalogs recommenders
2	8,663	0.8%	chough topping, crosslines request, expediency breakup, marksmanship centre, skirts hijack, loves mope, beautification risk, privacy psychotherapies, bounds genoa, treatment ascender
3-5	10,655	1.1%	gulfs handbooks, shores limbo, frigates minus, abuse spokes, limes mise, tear butts, dreary glumness, stylite eremite, blunders firms, manure hoeing
6-10	8,935	0.9%	prostitution commentary, minimal resists, stroller calls, opinion revel, perks peach, mommas japan, brook sheets, pedigree death, awed trade, jink bender
11-30	10,365	1%	brokenhearted advisor, vibration extremes, esteem ticket, plywood terminology, list levels, high carnations, microbial slaughter, egotism good, introductions occult, mystifying facial
31-99	7,806	0.8%	electronic retriever, old rattlers, shirts dancing, guitar desecrator, foreman children, sap esprit, toyland beach, komondor purebred, peeping caddis, imperative win
100-999	3,991	0.4%	blow rim, cheapest studio, zygomatic complex, lake shrinking, view pedestal, viper differential, seas kayaks, gallstones people, artwork mascot, story varieties
1,000-over	1,163	0.1%	comfortable beachfront, learning behavior, supplier enablement, important metal, factual questions, protease lactase, training resellers, particle precipitation, dentistry professionals, karmic patterns, login functionality

- **Language identification**: Some words listed in the DELA English dictionary used here are also common words in other languages. For example, both the words *de* and *ras* were listed as nouns in this dictionary, one as a singular and the second as a plural noun. In our simple juxtaposition method, the pair *ras de* was generated among the million pairs, and it was found on 76,144 Web pages. When the English language filter is turned on (not used in the counts here), this page count descends to 4,000. Using language tags to restrict counts to English pages may reduce the upper bound.
- **Not ignoring punctuation**: The index used to find the counts has removed punctuation, so that if two words are found next to each other with an intervening punctuation mark, these are considered as adjacent. For example, the snippet of text … *What are the symptoms of gallstones? People with gallstones don't know they have them until* … will match the query *gallstones people*.
- **Part of Speech tagging**: Similarly part of speech tagging can be used to identify that words such as *extracts* or *mention* are being used as verbs rather than nouns, for example in *It can decrease swelling, blood pressure and it extracts deposits left in lymph nodes*.
- **Shallow parsing**: If the inputs text were correctly chunked into noun phrases and verb phrases, then we would not recognize the phrase *training resellers* from *In the last several years I wrote the training materials and conducted the training resellers must attend to qualify for the …*
- **Two-word phrases**: Do not count subphrases appearing in a longer phrase. For example, the partial phrase *comfortable beachfront* is almost always part of a longer expression: *Comfortable beachfront home surrounded by old growth forests …*, *Stay at a comfortable beachfront hotel …*, etc.
- **Proper Name recognition**: We wouldn't recognize *assistants mike* from … *Assistants Mike Gallizzi, …*
- **Ignoring spam**: Spam is prevalent on the Web. Many Web pages exist only to post advertising, and they pilfer their content from elsewhere (newsfeeds, blog entries, Wikipedia). See the next section for more about spam.
- **Starting from the most common words**: Start from the most common words in each language, and build complete pair models for these words. We could also gather all the Web pages for words appearing 1,000

times or less and treat these pages completely, extracting all pairs found for each word.

These techniques will lead to better counts, and better approximations, but each has a processing cost that we could not incur here. A better-structured project on producing this estimate should implement most of them.

3.3. Caveats

This approach to mining the Web assumes that the Web contains raw material of worth. Unfortunately, it seems that in any human form of communication when someone is allowed to write, spam is a by-product. Over half of the e-mails sent these days is spam, and possibly half of the Web pages are also spam. Search engines implement procedures to recognize spam and to remove detected pages from their indexes. But every technique implemented to detect spam inspires new creativity in escaping detection.

(i) Bait and switch spam

Looking through generated pairs with low page counts, I checked a few to see where they came from. One combination, *eradicated compilation*, came from a Web page which contained the following paragraph, with only *compilation* appearing on the page:

> *The target of poker texas holdem game is to mix those 2 cards along with the five cards, that will at the end be situated before the dealer ('the board') in order to form the highest card combination. The winning hand can comprise any* ***compilation*** *of hole cards plus board-cards.*

But this text is not what the search engine sees or indexes. When the URL of the page is fetched from the Internet, a search engine finds the following text, in which I have underlined extra words not appearing on the final displayed page:

> *The gratuito target of poker texas holdem game is em to mix caesars those 2 cards along em with the five cards, stuff that will at mac the end be situated daytona before the dealer ('the home board') in original order to form the ich highest card combination. approximations The winning hand can use comprise any* ***eradicated compilation*** *of hole cards postmaster plus board-cards. Peruvianizes*

The extra text does not appear on the final page shown to the user because there is an obfuscated redirection implemented in javascript that is executed as soon as the page is loaded in a browser that shows the cleaned-up text.

The extraneous words appear to have been inserted randomly every three to ten words, and also randomly drawn from a list of English words. The purposes of the extra words are twofold: (i) increase the likelihood that the page is returned by a random query containing one of the added words, and (ii) primarily hide the fact that this cleaned text on the page has been stolen from some other site, since most search engines have implemented near-duplicate page detection. Near-duplicate detection will strip out side bars, menus, advertisements, and some numbers, and then compare the remaining text. If the same text is found on a higher ranking page, then this page is further deprecated, or removed from the index if from an already suspicious site.

The disadvantage of such spam, for language modelling purposes, is that odd combinations are created. Since they are random events, their frequencies remain low, but with over millions of spam pages, their cumulative counts can mask low-frequency but legitimate combinations.[13]

For such reasons, we cannot trust low-frequency counts as legitimate common usages.

(ii) Generated text

Another form of spam that creates unwanted phrases is found in the following example of generated text. The following Web page clearly wants to attract buyers of Zithromax (an Italian and Finnish brand name of a widely sold antibiotic, also called Azithromycin) to their Web page which directs to an online pharmacy store (of dubious authenticity). The page replicates structures of legitimate pages, with blog-like entries, but it seems to generate text using an algorithm that mixes text from a variety of sources, skipping from one text to another at stop words, as well as mixing in random words, words from a medical domain, and ensuring that sentence length falls within normal limits.

This page and five other spam pages attest the phrase *transnational distress* which probably originated as a typo of the medical term *translational distress*.

Abrade the ulterior.

And it's true, it has barely happened thousands of tetrahedron, so it must overly adapt to right itself on time otherwise I wouldn't be here, that's for sure. So perhaps you didn't know which drug they were helped or not you who amphoteric that here. I don't care. That's grouchy because we're taking

these drugs increase cyclo levels to a multi-year stoned course for low-dose accutane, which I think I am a troll over a year and you can't kill the zaire deader than dead, and extra drugs in TEST subjects.
Propionibacterium acnes vacillate you haven't clarification is the same guarantor that causes pimples (acne) in tyrant when our hormones start to spike unenlightened reason this salting sanity like the fickleness diluent.
I've done that with abx and was disappointed in the long term aspects. Try taking an antihistamine for a number of clientele who visit this f'ing newsgroup for 10 whodunit, or 5 utilization if Zithromax . Dawn I find ZITHROMAX scraped, not because of anything on his site about that. I'm not in transnational distress, but my liver enzymes are still closing up, a whole month after the end of May. Oh that ZITHROMAX may need to get him to educate it. First, I think the ZITHROMAX is that i have no reason to doubt the hell of your toxoid - why should I.

The random nature of such text assembly would throw noise into any language modelling. Consider calculating word association norms for the phrase *liver enyzymes*, for example, from such data.

(iii) Google counts

A final caveat is against using Google counts for any finer-grained statistical comparisons. As Jean Veronis[14], one of the first to detect this phenomenon, points out, the counts provided by Google are probably the result of extrapolating data from a smaller index. These counts depend on frequencies in this smaller index and the extrapolation is not reliable. I know that such extrapolation is not performed in the Exalead index used here.

4. Conclusion

Some of the potential applications of building a large-scale statistical description of word associations in the seminal paper *Word Association Norms, Mutual Information, and Lexicography* (Church & Hanks 1989) were to provide better language models for speech recognition and optical character recognition, as well as to help the lexicographer in making decisions on common usages of a word. Such techniques are also useful in determining proper translation in context.[15] The same language modelling applications (speech, OCR, translation) are still useful applications when we pass from word-based models to phrase-based models. Here we have tried to provide a first estimate of how many entries such

a phrase-based model should include if we want to cover all the commonly used concepts expressed as phrases. We estimate this to be about 200 million concepts that future computational lexicographers and linguists will have to find a way to model.

Notes

[1] Without ever really defining what I mean by 'concept,' operationally I will be using the word concept to mean something that is expressed using one or more words. When I use the word 'word,' I will mean a one-word concept.

[2] Cited in Brewer (2007: 102).

[3] Version 3.0 of WordNet contains 84,487 distinct words, plus 64,243 distinct multiword phrases. Cf. http://wordnet.princeton.edu/.

[4] Patrick believes that it is misleading to talk about a word having one sense or the other. He argues:

> The meaning potential of each word is made up of a number of components, which may be activated cognitively by other words in the context in which it is used. These cognitive components are linked in a network which provides the whole semantic base of the language, with enormous dynamic potential for saying new things and relating the unknown to the known.
>
> The target of 'disambiguation' presupposes competition among different components or sets of components. And sometimes this is true. But we also find that the different components coexist in a single use, and that different uses activate a kaleidoscope of different combinations of components. So rather than asking questions about disambiguation and sense discrimination ("Which sense does this word have in this text?"), a better sort of question would be "What is the unique contribution of this word to the meaning of this text?" (Hanks 2000: 214-215)

[5] See chapters 7 and 10 in Manning & Schütze (1999).

[6] Cf. Sag *et al*. (2002).

[7] In WordNet 3.0, there are 146,347 word entries with a noun tag and only 25,047 with a verb tag. Patrick has also called for the elimination of a number of unattested words from WordNet, such as the verbs *minify* and *desquamate* which would reduce this number of verbs (cf. Hanks 2004: 3).

[8] This depends, of course, on the ultimate application. Currently, for retrieving video, it is estimated that 'a few thousand semantic concepts could be sufficient to support high accuracy video retrieval systems' (Hauptmann *et al*. 2007: 633).

[9] Cf. http://www.askoxford.com/asktheexperts/faq/aboutenglish/numberwords.

[10] Which would mean some kind of infectious fermentation located on the end opposite to that on which the mouth is situated.

[11] The complete utterance from which it is drawn even has its own Wikipedia entry, at http://en.wikipedia.org/wiki/Colorless_green_ideas_sleep_furiously.

[12] One could easily imagine a parlour game built around these pairs.

[13] For example, the generated phrase *dusky chum* appears only once in our index, but from a legitimate textual source: *O'Tway's last recorded telephone call was to fellow Gibbon Martyn Johansson, to assure him that all was well, and that he was at that moment drinking several pints of cider in the company of a dusky chum of his, NKPWW Salve. However, according to the black box data recorder, he was actually sitting in the back of a Heathrow taxi, heading for Hampshire …*

[14] See his blog at http://aixtal.blogspot.com/2005/02/web-googles-missing-pages-mystery.html. We should remark that Mark Liberman also warned about using Google as a statistics machine for linguistic processing as early as 2003, see his blog at http://itre.cis.upenn.edu/~myl/languagelog/archives/000194.html. A more recent examination is Eu (2008).

[15] Cf. Jang *et al*. (1999).

References

A. Dictionaries

Hanks, P. (chief ed.)**, *et al.*** 1995. *Oxford English Reference Dictionary*. New York: Oxford University Press.

Hanks, P. (ed.)**, T. H. Long** (managing ed.)**, L. Urdang** (editorial director)**, *et al.*** 1979. *Collins Dictionary of the English Language*. (First edition.) London: William Collins Sons & Co. Ltd.

Hanks, P. (ed.) **& S. Potter** (editorial consultant). 1971. *Encyclopedic World Dictionary*. (First edition.) London: Paul Hamlyn Publishers.

B. Other literature

Brewer, C. 2007. *Treasure-House of the Language: The Living OED*. New Haven, CT: Yale University Press.

Church, K. W. & P. Hanks. 1989. 'Word Association Norms, Mutual Information, and Lexicography' in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 26-29 June 1989*. Vancouver: University of British Columbia, 76–83.

Eu, J. 2008. 'Testing Search Engine Frequencies: Patterns of Inconsistency'. *Corpus Linguistics and Linguistic Theory* 4.2: 177–207.

Good, B. M. & J. T. Tennis. 2009. 'Term Based Comparison Metrics for Controlled and Uncontrolled Indexing Languages'. *Information Research* 14.1: paper 395. Online at http://InformationR.net/ir/14-1/paper395.html.

Guthrie, L., J. Pustejovsky, Y. Wilks & B. M. Slator. 1996. 'The Role of Lexicons in Natural Language Processing'. *Communications of the ACM* 39.1: 63–72.

Hanks, P. 2000. 'Do Word Meanings Exist?' *Computers and the Humanities* 34.1-2: 205–215.

Hanks, P. 2004. 'WordNet: What is to be Done?' *Panel Presentation at the Second International WordNet Conference (GWC 2004), Brno, Czech Republic, January 20-23, 2004*. Brno: Faculty of Informatics, Masaryk University, 1–6. Online at http://www.fi.muni.cz/gwc2004/pres/panel/Hanks/hanks-panel.pdf.

Hanks, P. 2005. 'Johnson and Modern Lexicography'. *International Journal of Lexicography* 18.2: 243–266.

Hauptmann, A., R. Yan & W.-H. Lin. 2007. 'How Many High-level Concepts will Fill the Semantic Gap in Video Retrieval?' in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval* (CIVR 2007). Amsterdam: University of Amsterdam, 627–634.

Jang, M.-G., S.-H. Myaeng & S.-Y. Park. 1999. 'Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting' in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, MD: University of Maryland, 223–229.

Manning, C. D. & H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Margolis E. & S. Laurence. 2007. 'The Ontology of Concepts – Abstract Objects or Mental Representations?' *Noûs* 41.4: 561–593.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross & K. J. Miller. 1990. 'Introduction to WordNet: An On-line Lexical Database'. *International Journal of Lexicography* 3.4: 235–244.

Sag, I. A., T. Baldwin, F. Bond, A. A. Copestake & D. Flickinger. 2002. 'Multiword Expressions: A Pain in the Neck for NLP' in *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (Lecture Notes in Computer Science 2276). Berlin: Springer, 1–15.